(1) (a) since y is a one-hot vector with 1 at word o and 0 elsewhere,

$$-\left[y^T \log(\hat{y})\right] = -\left[\sum_{w \in Vocab} y_w \log(\hat{y}_w)\right]$$

$$= -\left[\sum_{j \neq o} (0)(\log(\hat{y}_j)) + (1)\log(\hat{y}_o)\right]$$

$$= -\log(\hat{y}_o)$$

(1) (b) $J_{\text{naive softmax}}(v_c, o, U) = -\log\left[\dfrac{e^{u_o^T v_c}}{\sum_w e^{u_w^T v_c}}\right]$

$$\frac{\partial J}{\partial v_c} = \frac{\partial J}{\partial \hat{y}_o} \cdot \frac{\partial \hat{y}_o}{\partial v_c} = -\frac{1}{\hat{y}_o}\left[\frac{\sum_w e^{u_w^T v_c}(u_o e^{u_o^T v_c}) - e^{u_o^T v_c}\left(\sum_k v_c e^{u_k^T v_c}\right)}{\left(\sum_v e^{u_v^T v_c}\right)^2}\right]$$

$$= \frac{-1}{\hat{y}_o}\left[\frac{\sum e^{u_w^T v_c}(u_o e^{u_o^T v_c}) - e^{u_o^T v_c}\left(\sum_k u_k e^{u_k^T v_c}\right)}{\left(\sum_v e^{u_v^T v_c}\right)^2}\right]$$

$$= \frac{-1}{\hat{y}_o}\left[(1)(u_o \hat{y}_o) - (\hat{y}_o)\left(\sum_k u_k \hat{y}_k\right)\right]$$

$$= -u_o + \sum_k u_k \hat{y}_k = -\sum_i y_i u_i + \sum_k u_k \hat{y}_k$$

$$= -U^T y + U^T \hat{y} = U^T(\hat{y} - y)$$

(1) (c) $\frac{\partial J}{\partial u_w} = \frac{\partial J}{\partial \hat{y}_w} \cdot \frac{\partial \hat{y}_w}{\partial u_w}$

Let $w = o$, $\left(-\frac{1}{\hat{y}_o}\right)\left[\frac{\left(\sum_w e^{u_w^T v_c}\right)\left(v_c e^{u_o^T v_c}\right) - \left(e^{u_o^T v_c}\right)\left(v_c \sum_w e^{u_w^T v_c}\right)}{\left(\sum_w e^{u_w^T v_c}\right)^2}\right]$

$$= \frac{-1}{\hat{y}_o}\left[\frac{\left(\sum_w e^{u_w^T v_c}\right)\left(v_c e^{u_o^T v_c}\right) - \left(e^{u_o^T v_c}\right)\left(v_c e^{u_o^T v_c}\right)}{\left(\sum_w e^{u_w^T v_c}\right)^2}\right]$$

$$= \frac{-1}{\hat{y}_o}\left[(1)(v_c)(\hat{y}_o) - v_c(\hat{y}_o)^2\right] = v_c(\hat{y}_o - 1)$$

Let $w \neq o$, $w = z$, $\frac{-1}{\hat{y}_o}\left[\frac{\left(\sum_w e^{u_w^T v_c}\right)(0) - \left(e^{u_z^T v_c}\right)\left(v_c e^{u_o^T v_c}\right)}{\left(\sum_w e^{u_w^T v_c}\right)^2}\right]$

$$= -\frac{1}{\hat{y}_o}\left[-\hat{y}_z v_c \hat{y}_o\right] = \hat{y}_z v_c = v_c(\hat{y}_z - \cancel{y_z})$$

$$\therefore \frac{\partial J}{\partial U} = \underset{N\times 1}{(\hat{y} - y)}\ \underset{1\times D}{v_c^T}$$

(1) (d) $\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}$

$$\sigma'(x) = \frac{(e^x+1)(e^x) - (e^x)(e^x)}{(e^x+1)^2}$$

$$= \sigma(x) - \sigma(x)^2 = \sigma(x)(1-\sigma(x))$$

(1) (e) $$J_{neg\text{-}sample}(v_c, o, U) = -\log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} \log(\sigma(-u_k^T v_c))$$

$$\frac{\partial J}{\partial v_c} = \frac{-1}{\cancel{\sigma(u_o^T v_c)}} \cdot \left[\cancel{\sigma(u_o^T v_c)}\right] \cdot \left[1-\sigma(u_o^T v_c)\right](u_o) - \sum_{k=1}^{K} \frac{1}{\cancel{\sigma(-u_k^T v_c)}} \cdot \cancel{\sigma(-u_k^T v_c)} \cdot \left[1-\sigma(-u_k^T v_c)\right](-u_k)$$

$$= -u_o\left(1-\sigma(u_o^T v_c)\right) + \sum_{k=1}^{K} u_k \left(1-\sigma(-u_k^T v_c)\right)$$

$$\frac{\partial J}{\partial u_o} = \frac{-1}{\sigma(u_o^T v_c)} \cdot \sigma(u_o^T v_c) \cdot \left(1-\sigma(u_o^T v_c)\right) v_c$$

$$= -v_c\left[1-\sigma(u_o^T v_c)\right]$$

$$\frac{\partial J}{\partial u_k} = \frac{\cancel{-}1}{\cancel{\sigma(-u_k^T v_c)}} \cdot \cancel{\sigma(-u_k^T v_c)} \cdot \left(1-\sigma(-u_k^T v_c)\right)(\cancel{-}v_c)$$

$$= v_c\left(1-\sigma(-u_k^T v_c)\right)$$

use for-loop for neg samples. Possible Duplicates

Negative sampling is more efficient than naive softmax. When computing $\frac{\partial J}{\partial v_c}$, naive softmax uses the entire word matrix U, whereas negative sampling uses only the target outside word vector, u_o and K negative samples $u_1, \ldots, u_K$ to compute gradients.

(1) (f) $J_{skip\text{-}gram}(v_c, w_{t-m}, \ldots, w_{t+m}, U)$

$$= \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} J(v_c, w_{t+j}, U)$$

(i) $$\frac{\partial J_{SG}}{\partial U} = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \frac{\partial J(v_c, w_{t+j}, U)}{\partial U}$$

(ii) $$\frac{\partial J_{SG}}{\partial v_c} = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \frac{\partial J(v_c, w_{t+j}, U)}{\partial v_c}$$

(iii) $$\frac{\partial J_{SG}}{\partial v_w} = 0$$

$w \neq c$